

FORECASTING SALES, FOR THE SMALL
READY-MIX CONCRETE COMPANY

By

DOUGLAS LATHEL JOHNSON

Bachelor of Science

Oklahoma State University

Stillwater, Oklahoma

1963

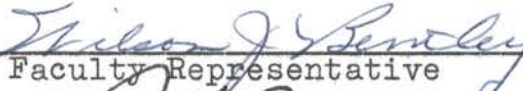
Submitted to the Faculty of the Graduate School
of the Oklahoma State University
in partial fulfillment of
the requirements for
the degree of
MASTER OF SCIENCE
May, 1964

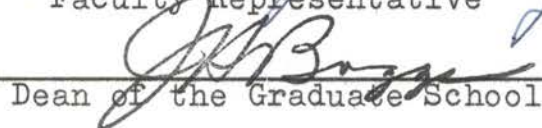
JAN 6 1965

FORECASTING SALES FOR THE SMALL
READY-MIX CONCRETE COMPANY

Thesis Approved:


Thesis Adviser


Faculty Representative


Dean of the Graduate School

569777

PREFACE

Forecasting of demand is a subject often encountered in the Industrial Engineering curriculum. Many problems in planning and inventory control base their solution on the ability to accurately forecast future needs and conditions.

Several attempts have been made at establishing inventory policies that will serve the Ready-Mix Concrete industry. However, the success of these efforts has been limited by the difficulties inherent in forecasting future demand for an industry that experiences a highly seasonal and generally volatile sales pattern.

It was the desire to develop a method that could be adequately used for forecasting under the above conditions that has generated this thesis.

I am deeply indebted to the Ideal Cement Company for the fellowship that permitted the undertaking of this project. I am further indebted to Dr. Paul E. Torgersen for his continued guidance and assistance throughout the course of this study.

I would like to thank the Murphy and Perkins Concrete Company, of Oklahoma City, Oklahoma, and the Morrow Service Company, of Perry, Oklahoma, for supplying the sales data

used in this study.

It is with deepest gratitude that I thank all those whose friendship and effort have made possible not only this study but my entire graduate program.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
Purpose of the Study	4
II. SALES DATA	7
Method of Analysis	8
III. THE FORECASTING EQUATION	20
IV. SUMMARY AND CONCLUSIONS	30
Extensions of the Study	33
APPENDIXES - INTRODUCTION	38
PROBLEM I	40
PROBLEM II	44
PROBLEM III	48
PROBLEM IV	54

LIST OF TABLES

Table	Page
I. Variable Lead Values	14
II. Representative Correlation Matrix	16
III. Selected Intercorrelations	18
IV. Forecasts for 1962 and 1963	23

CHAPTER I

INTRODUCTION

Forecasting future sales or work requirements is a problem faced by every company doing business. In a large organization, forecasting may take the form of a quite complex analysis of the many factors considered to affect the product or products of the company. In the smaller company, forecasting may consist of a simple intuitive extension of past experience into the future. In some industries, past experience is a strong indicator of activity that may be expected in the future. However, this type of forecasting can only be successful if the conditions affecting the particular products of the company remain relatively fixed, or if changing, change uniformly.

Whatever the problems encountered in determining future sales or material requirements, it is very necessary for management to be able to plan. In order to plan for inventory requirements, production schedules, labor requirements, financial loads and allocations, and future growth and expansion, company management must have some indication of the volume of demand to be expected for their product. If an extremely good year is anticipated, the company must

make plans for hiring additional labor, enlarging inventories of required materials, purchasing additional equipment, and financing these extended operations. If an extremely poor year is expected, plans must be made for reductions and reallocations of these same resources. If business is expected to continue at the same pace experienced in the past year or years, management must again be aware of this condition, so that ill-considered or ill-timed expansions, reductions, or reallocations may be postponed or prevented.

Accurate forecasting is, in many cases, difficult or impossible. In many industries, the factors affecting company sales cannot be determined, or, if determined, cannot be accurately measured. In large organizations, producing many products, the problems caused by unexpected fluctuations in one product can be absorbed by gains on other products. However, in a smaller company, with only one or a very few products, wide or unexpected fluctuations in sales can be quite damaging. It is, however, in these very companies that forecasting is the most difficult. In a small company, the necessary manpower required for developing an adequate forecast is not available, or if available, does not have adequate training and experience. In many cases, it is extremely difficult to determine the factors responsible for fluctuations in sales. How, then, can sales for a company of this type be forecast?

There are numerous methods available for forecasting. Some of these are quite simple and easily applied. Many are based entirely on extrapolation of historic data to future periods. However, due to both their simplicity and the generality of their assumptions, these methods are often inadequate for use in planning future operations.

The Ready-Mix concrete industry is one in which these conditions are found. The small Ready-Mix concrete company experiences an extremely volatile sales pattern. Part of the variation can obviously be explained by seasonal fluctuations, but there remains a large amount of variation whose source is not obvious. Due to the relatively small nature of most of these companies, much or possibly most of this variation might be in response to the vagaries of the consumer. For instance, in a small Ready-Mix concrete company, a single order of any appreciable magnitude can be sufficient to throw the sales pattern from one extreme to another. A large company in this industry does not experience this, for their sales volume is high enough that the influences from individual orders tend to average out over the sales period. However, the small company is often aware of a particularly large order for some time before it must be delivered. Thus, in planning future sales, these unusual orders are provided for in excess of the normally expected sales for that period. Thus, the primary question is, "What are the normally expected sales for some future period?"

Purpose of the Study


The purpose of this study is to develop some means of accurately forecasting sales for the small Ready-Mix concrete company. Past work in developing inventory policies for these companies has pointed to the need for an adequate prediction of demand in order to more successfully apply the policies constructed.

In order to develop a predictor to satisfy this need, it was decided to attempt to develop a linear regression equation based on variables, in time series form, that appeared to be logically related to the volume of concrete sales. These variables were not expected to relate in a true cause and effect manner with the experienced concrete sales, but rather to be factors whose actual volume or production might be used to indicate the magnitude of sales volume to be expected for concrete in later sales periods. These variables, referred to as leading indicators in the field of economics, were to be tested and either used or eliminated on the basis of their effect on the linear regression equation developed. Those variables that in combination yielded a minimum prediction error would then be used as the mechanism for predicting future sales.

Linear regression is a very powerful tool, but must be used with care. Haphazard use of regression can result in misleading results. It is often possible to find highly

correlated relationships between variables that have no logical connection. It is, of course, possible that the indicated relation is one that does truly exist, but is not intuitively or theoretically obvious. It is more likely, however, that the relation is merely a chance occurrence, and one that may change or fail to continue at any time in the future. The problems that might result from forecasts based on variables of this kind are obvious. If extensive capital planning is based on this type of forecast, large losses might be incurred if the chance relationship on which the forecast was based ceased to hold, causing the forecast to be badly in error. To avoid this pitfall, variables must be selected on a logical basis. Further, any extreme or unusual forecasts, or any unexpected or unusually close relationships should be very carefully examined. If any indication is found that the relationship is unrealistic, extreme care should be exercised in using the forecast values thus derived. However, if care and intelligence is used in the initial selection of variables, problems of this type should not be encountered.

It is the purpose of this study, then, to select a fairly large number of variables, logically related to sales in the concrete industry, and through manipulation in linear regression equations, determine that combination of variables best describing the sales pattern found in particular Ready-Mix concrete companies. The computer



programs and methods of analysis developed for studying a particular company should be easily applicable to other companies in this industry. The exact methods of this study will be discussed in detail in the following chapters, however, it should be recognized that these techniques can be applied in any case in which the analyst is able to logically select variables that either indicate or cause the effect of interest.

CHAPTER II

SALES DATA

In order to conduct this study, sales data was obtained from two Ready-Mix Concrete companies. These companies are of quite different size, and are in two different locations. The data was received in the form of monthly sales totals, without further breakdown. These totals include all sales made during the indicated month, and no allowance is made for orders of unusual size or type. Thus, fluctuations from this source will not be accounted for in the final predictor, so that allowance must be made prior to application and use of the forecast. This must be done on the basis of known orders on hand at the time of the forecast. Since these orders are small in number, their occurrence appears almost random in nature. If annual or biannual sales totals were used, it is possible that the total effect of these orders might be accounted for adequately. However, the number of unusual orders occurring in a single month is so small that it is impossible to predict.

Method of Analysis

The method of analysis used in this study follows a systematic procedure of selection and elimination of variables, until a regression equation is derived, involving the best combination of variables it is possible to obtain from the independent variables initially selected. The steps in this analysis will be discussed below in the order in which they occur.

First, the concrete industry, in general, was carefully studied in order to determine those factors that might be expected to most prominently affect production and sales. For instance, such factors as the general state of the economy, volume and trends in the construction industry, and income and available capital might be expected to either affect or indicate the volume of sales experienced by the industry. Thus, it is necessary to select time series that adequately represent these general economic conditions. Since sales volume in the concrete industry can be expected to respond in some manner to these factors, the sales of a particular company in this industry can be expected to respond in a like manner to the fluctuations of variables representing these segments of the economy.

For purposes of this study, 16 variables were selected from various sources. These variables, and the reasons for their selection are given below. Since both companies

used in this study were Oklahoma companies, several of the following indicators are state oriented. This was done to give weight to regional factors that might be expected to affect sales.

Agriculture production and total personal income for Oklahoma were chosen as state measures of available capital and as a measure of economic conditions in the State (1). It was felt that much of the concrete sold by the smaller Ready-Mix companies would be to individuals. Since, in the smaller towns, much of these sales might be assumed to be agriculturally oriented, the income from agriculture production was included as an indicator for this segment of the economy. The personal income series was expected to indicate the general state of private finances; the assumption being that the higher the general level of income, the greater would be the amount of money spent for property improvements and construction.

Freight car loadings is a series generally believed to indicate the level of business activity being experienced (1). For this reason, the total monthly car loadings for Oklahoma was included as a further indicator of the general economic level in the State.

The total dollar value of Oklahoma construction, by month, was included as an indicator of construction activity in the State (1). This series is primarily based on construction contracts awarded, thus it is expected to indicate any change in the activity of the construction

industry. Since these are contracts for future construction, any change in the level of this series should indicate a corresponding change in the amount of concrete sold in future periods.

Time series data for the production of brick, concrete, and plaster products, construction steel, construction materials, metal materials, cement, and lumber was included as a measure of national construction activity (2). It was expected that the sale of Ready-Mix concrete would generally follow the movements of one or more of these products, since each is used in construction, and their production is gaged to meet either known or expected demand in present or future periods. Materials, such as brick, lumber, etc., must be ordered prior to actual construction, so that it will be available when needed. Ready-Mix concrete is made for immediate use on the construction site. Thus, the production of other construction materials for any given activity should proceed, by some consistent time interval, the production of concrete for those same activities.

The Federal Reserve Board index of total industrial production was included as a measure of the state of the national economy (2). This series was expected to indicate major trends in the economy which might affect the level of activity in the construction industry, and thus, the level of activity in the Ready-Mix concrete industry.

Variables representing weather conditions were also

considered initially. It was found extremely difficult to locate monthly data for either temperature or rainfall. Consideration of the daily and weekly data available indicated that monthly totals, or averages, would tend to conceal the effects of interest. If forecasting was being done on a daily, or possibly weekly basis, rainfall and average temperature would be expected to indicate the amount of working time lost due to weather conditions. However, total rainfall per month gives no indication of the number of days during which rain fell, for two days of heavy rains might show the same total as 15 days of light or medium rains. In the same manner, average temperature gives no indication of the number of days in the month that were above or below a certain temperature. For these reasons, then, variables relating to weather were not included.

Time series data was also considered for the book value of manufacturers' inventories, the Federal Reserve Board building cost index, the national production of Portland cement, and the level of crude oil production in Oklahoma (3), (2), (4), (1). These series were expected to indicate various factors in the economy, but were later eliminated due to either insufficient or inaccurate data, an obvious lack of relationship, or a nebulous logical relationship.

The elimination of these four variables reduced the number of real variables under consideration to twelve.

In addition to these twelve real variables, four artificial variables were included. The first three of these were variables designed to account for quarterly seasonal variation in the sales data. Each variable had a value of one for each month of the quarter to be adjusted, and zero for all other months. By adjusting the first three quarters of each year, adjustment of the fourth quarter was automatically insured. The fourth of these artificial variables consisted of numbers in sequence from one to the total number of observations. This variable was included as an adjustment for any upward trend in the data not explained by the other variables. This type of variable gives weight to the growth rate of the particular company of interest, plus any other unexplained trends encountered.

The second step in the analysis consisted of plotting the selected variables on a common time scale. This allowed careful visual consideration of trend, cycles, degree of stability, and apparent relationship with the company sales data. Since these variables, to be of use as predictors, must lead the sale of concrete by one or more months, it was necessary to determine those lead intervals indicating the best correspondence between the various series and the sales data. By placing each variable over the plots of company data, and aligning the time axis, it was possible to study apparent likenesses. Each variable was compared at a number of different times, by

shifting the time axis from one to fifteen months ahead of company sales. This procedure made it possible to select several intervals for further comparison.

The procedure used in determining the appropriate number of periods lead, was to shift the time axis until the apparent coincidence of the two series was best. The number of periods the axis had been shifted was then designated as the best lead. Values on either side of the best were selected to yield a range of possible lead values for consideration in later analysis. Three to five lead values were selected for each variable. From these values for each of the twelve variables, six likely combinations of lead values were developed. These combinations involved the best lead value for each variable, the next best for each variable, and selected combinations of best, second best, etc. Table I shows the lead values, in months, as developed from this step in the analysis.

This procedure for selecting leads was chosen as one method of selecting the most likely leads for further analysis without testing all possible combinations. Several values were selected for each variable, since the comparison was visual, and accuracy was not insured. It was felt, that by testing values within a two to three month interval, about the apparent best lead, the true best lead would be more certain to be included.

The third step in the analysis was to search for intercorrelations among the variables. By eliminating any

TABLE I
VARIABLE LEAD VALUES

	Variable No.											
	1	2	3	4	5	6	7	8	9	10	11	12
Lead Set 1	1	1	1	1	1	1	2	1	7	1	3	1
2	2	2	2	2	2	2	10	2	8	2	4	2
3	3	3	3	3	1	3	11	1	9	0	5	3
4	0	2	2	1	1	1	11	1	8	0	4	1
5	2	4	3	2	2	3	12	2	10	2	5	2
6	3	4	4	3	2	3	2	2	7	2	5	3

Variable No.	Variable Name
1	Agriculture Production (Okla.) ✓
2	Total Personal Income (Okla.) ✓
3	Freight Car Loadings (Okla.)
4	Oklahoma Construction
5	Brick
6	Concrete and Plaster Products

Variable No.	Variable Name
7	Construction Steel
8	Construction Materials
9	Metal Materials ✓
10	Lumber ✓
11	Total Industrial Production ✓
12	Cement

TABLE II
REPRESENTATIVE CORRELATION MATRIX

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.0	.64			.74	.73		.72		.62	.66	.79
2		1.0	.79	.75	.82	.95	.66	.94	.78	.79	.96	.84
3			1.0	.86		.75		.69			.71	.64
4				1.0		.75		.66			.66	.64
5					1.0	.90	.66	.94	.77	.92	.88	.94
6						1.0	.62	.97	.75	.82	.94	.93
7							1.0	.70	.76	.76	.72	
8								1.0	.83	.90	.97	.92
9									1.0	.82	.85	.68
10										1.0	.88	.82
11											1.0	.85
12												1.0

visual correspondence to the sales data, it was included. It was felt that the indicated correlation was not serious enough for the elimination of either variable.

This analysis gave rise to several relatively high correlations between some pairs of variables. Some of these are shown in Table III, strictly as a matter of possible interest. Some of these might be expected, others, however, might be of possible use in other studies.

The fourth step in the analysis is the determination of the best possible regression equation as determined by an evaluation of all possible regressions on the selected variables. The four independent variables selected by the correlation analysis and the four artificial variables previously discussed were used in this regression analysis. These eight variables were regressed against the sales data for each of the two companies involved in this study. Under program control, all possible regression equations that could be developed from combinations of these eight variables were tested. This computer program is described in Appendix III. Due to the data format required by the program, a new data deck involving only the eight selected variables had to be developed. This too was done under program control, as shown in Appendix I. This program is included only in the interest of a complete description of the process, for it is simply an input/output device used in the interest of saving time and effort. Since ten years data was used for each variable, the time saved

TABLE III
SELECTED INTERCORRELATIONS

Variable	Leads	Variable	Correlation
Total Personal Income	3 mo.	Concrete and Plaster Products	.95
"	3 mo.	Construction Materials	.94
"	0 mo.	Total Industrial Production	.95
Brick Production	0 mo.	Concrete and Plaster Products	.90
"		Lumber Production	.91
Concrete and Plaster Products	1 mo.	Total Industrial Production	.94
Construction Materials	-3 mo.	"	.97

was quite significant.

In the interest of time, the program used to evaluate the regression equations computed only the coefficient of determination, and the residual sum of squares for each equation. The value of the equations were developed on the basis of these values.

All regressions were evaluated for each set of lead values, in order to determine that set of leads giving the best equation. The result of this step in the analysis, then, was a list of the variables that should be used for the best linear regression equation it was possible to obtain from the variables included for enumeration. These variables then served as input to an IBM library program which developed the regression coefficients for the equation. A copy of this program is included in Appendix IV.

The regression equation obtained from this analysis is the equation to be used for forecasting sales. The standard error of estimate was computed, and forecasts made for 1962, and the first three months of 1963. The actual equation, the error, and the forecasts are shown in the next chapter. A discussion of problems encountered in this analysis, and suggested improvements are presented in Chapter IV.

CHAPTER III

THE FORECASTING EQUATION

The equation developed from the previous analysis for the first company is

$$Y = 71,432 + .0441X_1 + .2344X_2 + 7,284X_3 - 399.8X_4 \\ - 9,533X_5 + 2,904X_6 + 6,735X_7 + 221.8X_8$$

where

Y is the dependent variable representing
expected sales

X_1 is agriculture production for Oklahoma

X_2 is Oklahoma construction

X_3 is construction steel production

X_4 is cement production

X_5 , X_6 , X_7 are artificial variables for
averaging seasonal effect

and X_8 is the artificial variable representing
unexplained trend.

The lead values appropriate to this equation are as follows:

3 months for X_1

3 months for X_2

2 months for X_3

3 months for X_4

and 0 months for X_5 , X_6 , X_7 , and X_8 .

A lead of 3 months indicates that if one were forecasting for January, 1963, the value of variable X_1 for October, 1962 would be used in the equation.

The coefficient of determination for this equation is .4704, with a residual sum of squares of 2.646×10^{10} . The coefficient of determination indicates that approximately 47% of the variance experienced in Ready-Mix concrete sales is explained by the eight variables included in the equation. The coefficient of multiple correlation for this equation is .6858, and the standard error of estimate as computed from the formula

$$E = \sqrt{\frac{\text{Residual Sum of Squares}}{n - 1}},$$

where n is the number of observations on the dependent variable, is \$7,600. This means that the expected difference between the actual and forecast sales is 7,600 dollars. Thus, if the differences between actual and forecast sales were averaged over a long period of time, the computed error would be this average. This value represents an error of approximately 20% of the average period sales.

It must be realized that when dealing with a sales pattern showing wide fluctuations, the forecast error should also be expected to fluctuate widely. In some instances, the error may be so small that it may be considered negligible. In others, however, the error may be almost as large as the sales for that period. Thus, since the average error is so large, the incidence of extreme error values would probably be frequent. For this reason, forecasting with an equation that exhibits an error of this magnitude is extremely unreliable. Table IV shows values forecast from this equation compared with the actual sales experienced, for 1962 and the first three months of 1963.

As explained above, and as should be expected from consideration of the magnitude of the error of estimate, and the small amount of variance explained by the equation, the forecast values show large deviations from actual sales. It is not difficult to imagine that a person, with some years experience in the industry, could intuitively forecast sales more accurately. However, consideration of Table IV indicates some possible sources of error that might lead to a more accurate forecast if corrected.

Consideration of the quarterly, semi-annual, and annual totals shown indicates that agreement between forecast and actual sales tends to be better when considered over longer periods of time. Quarterly totals for the second and third quarters for 1962 show very close

TABLE IV
FORECASTS FOR 1962 AND 1963

Month	Actual Sales (Coded)	Forecast Sales	Quarterly Totals	6 mo. Totals
Jan.	19,557	32,504		
Feb.	33,917	34,879		
Mar.	34,800	36,989	A: 88,275 F: 102,373	
Apr.	52,715	46,042		
May	46,985	46,795		
June	37,209	46,790	A: 136,810 F: 139,629	A: 225,085 F: 242,002
Jul.	46,015	44,056		
Aug.	47,820	41,503		
Sept.	34,884	43,281	A: 128,720 F: 128,840	
Oct.	44,394	39,061		
Nov.	43,769	37,201		
Dec.	37,218	38,291	A: 125,382 F: 114,553	A: 254,102 F: 243,393

Jan.	23,736	35,459	
Feb.	46,449	38,432	
Mar.	64,922	48,649	A: 135,107 F: 122,541

agreement, while the totals for the first and last quarters show an apparently smaller error than did the monthly totals. Although this same effect might be observed in many situations, and considered merely an averaging of random errors, there appears to be a better explanation in this particular situation. It will be remembered that the seasonal affect was averaged for each quarter, with forecast sales receiving only this quarterly adjustment for seasonal affect. It is felt that, due to the extremely seasonal nature of this industry, a quarterly adjustment was not sensitive enough to completely eliminate seasonal variance. By developing a seasonal index for company sales, by month, and using only seasonally adjusted sales figures in the analysis, it is believed that much better accuracy of forecast could be obtained. This method, of course, would eliminate the three variables included to adjust for seasonal variation. Thus, more real variables could be included within the limits of the computer program, perhaps explaining still more of the variation found in the sales data.

A further adjustment that might help to improve the accuracy of the forecast is to adjust the constant term in the equation for the average value of large jobs contracted for future periods. It has been previously mentioned that a portion of the variation in monthly sales totals might be explained by the occurrence of a small number of large orders. Since this number is quite small, it is almost

impossible to predict. However, by carefully analyzing past sales, it would be possible to determine the average value of jobs of this type. Orders of this nature should be fairly easily distinguished, since they will be many times the size of orders normally experienced. This average value should then be deducted from the constant term in the equation. The modified equation would then forecast normal or expected orders without giving weight to large or unusual orders. Since orders of this magnitude would normally be placed some time before delivery, it would then be possible to simply add to the forecast the actual value of these unusual orders. Under this procedure, normal sales should be much more accurately forecast, since the occurrence or non-occurrence of unusually large orders is no longer a factor. By adding the actual value of those orders known to be due during the next period, the accuracy of the forecast does not suffer, for unusual sales are added in the exact amount in which they will occur. It is, of course, possible that large orders might be placed without warning, but situations of this nature are present in any free market, and must be met or ignored as prescribed by company policy, and available production capacity.

The above adjustment may be relatively easily made by the user, and tested against the last year's sales. However, if the equation is adjusted in this manner, the standard error of estimate computed above is no longer

valid. Furthermore, it is impossible to compute a new error value, since the adjustment was made to the equation rather than to the data. If the new equation showed fairly close agreement with actual sales values throughout the past year it would probably be safe to use, if used cautiously.

If the actual sales data were modified by eliminating orders of unusual magnitude, and a new equation developed, it would be possible to compute a new error of estimate. This, of course, would be the safest method, and would be relatively easy to accomplish once the basic sales data had been modified. Modification of the sales data, however, might be quite difficult and time consuming, depending on the amount and type of historical data available.

If the data was to be modified in this manner, a seasonal index should also be developed, and applied at the same time. The equation developed from data of this type should be much more accurate, and, therefore, much more useful.

One further point regarding use of any equation of this type should be made at this time. A linear regression equation represents the apparent relationship among those variables included in the analysis. It does not, however, insure that this relationship is exact, or that it will continue to hold indefinitely. It can only be said, that the relationship indicated is one that has occurred in the past, and will continue to hold in the

future only so long as conditions underlying the relation continue unchanged.

The variables used in this study appear to be relatively stable, and barring violent upset in the economy generally, or in the industries giving rise to the series involved, should continue in essentially the same pattern. If at any time, however, forecasts are badly in error, or consecutive forecasts tend to be in error in the same direction, continued use of the same equation should be made with care. If either of these conditions appears, careful study should be made of the variables in use, and of the economy in general, in an effort to determine the cause of the deviation. If it is found that a change has occurred in the nature of either the economy, or one or more of the variables in the equation, a new equation should be developed using either the same variables, or new variables in place of those that have changed. In any case, once deviation from the normal pattern of accuracy is experienced, great care should be exercised in further use of the equation, until the reason for deviation has been discovered and corrected.

The past few pages have been devoted to an explanation of the equation developed for the first company analyzed. The analysis for this company was complete, and comments have been made concerning modification and use of the equation developed. However, data from a second

company was included in this study. A complete analysis was not made for this company for several reasons. The most important of these reasons was that preliminary analysis indicated the best regression equation that could be developed for this company would explain only 23% of the variance in the sales data. Forecasts from this equation would most likely have an error greater than the average monthly sales experienced by the company. Several reasons for this result seem apparent. The sales for this company were much smaller than for the other. For this reason, the problem almost becomes one of predicting the actions of individual customers, rather than the average response of a large population of customers. This is a more exaggerated case of the occurrence of unusually large orders mentioned in connection with the previous analysis. In this case, however, orders from a few customers can radically affect the total sales for the month. Thus, adequate adjustment is extremely difficult.

It was further expected that the affect of the seasonal factor would again contribute significantly to the forecast error. Since this was felt to be a major factor in the reduced accuracy of forecasts for the first company, further analysis seemed fruitless.

Due to the small amount of variance explained by the variables included in the analysis, it seems likely that these variables do not adequately represent the factors affecting the sales of this company.

These statements are not intended to indicate that it is impossible to forecast sales for this company, but are intended simply to indicate the difficulties encountered in the analysis of sales for a company of this size. Modification of the sales data for seasonal effect and unusual orders, as suggested in the previous analysis, and inclusion of variables more representative of the factors affecting the sales of small companies, should lead to a usable and reasonably accurate forecasting equation. Modification of the data in this manner, although not impossible, will be quite difficult, since it seems likely that a fairly large portion of sales might be attributable to relatively substantial orders from a few individual customers.

Thus, although the method of analysis is appropriate, considerable study and adjustment of sales data must be accomplished before a useful sales predictor can be derived.

CHAPTER IV

SUMMARY AND CONCLUSIONS

The method of analysis developed in this thesis consisted of the following four distinct steps:

1. The initial selection of time series variables considered to represent factors affecting sales volume, or considered to exhibit some logically sound relationship with sales volume.
2. The determination, by graphical techniques or any other appropriate method, of the combination or series of combinations of variable leads deemed most likely to represent the best correspondence between variables. It must be remembered that no matter how close a relationship between sales and some variable, the variable is not useful as a predictor if it does not lead sales by one or more periods.
3. The careful elimination of intercorrelation between variables used as predictors.
4. The development of the best possible regression equation from those variables remaining after the preceding analysis.

In studies of this type, it is normally recommended that the residual values remaining after elimination of known sources of variation be tested for autocorrelation. Evidence of the presence of autocorrelation should lead to adjustment of the error of estimate by one of the equations shown in Ezekial and Fox (5). This was not done in the present study, since the suggested modifications to the derived regression equation will greatly change the nature of the residual quantities. Detection of the presence or absence of autocorrelation at this point would be almost entirely meaningless.

The method developed in this thesis seems quite powerful, if carefully applied to accurate and reliable sales data. The problems encountered with the data used in the sample study should be remembered when applying this technique. One of the strongest features of this method is that it is primarily one of objective selection of appropriate variables from a continually refined population of those deemed applicable. It is generally believed that one of the most difficult problems to surmount in forecasting with a linear regression equation is determining accurately those variables most appropriate for inclusion in the equation. Use of this method allows one to begin with almost any number of variables thought generally applicable, and successively reduce their number until those remaining in the final equation yield the best fit to the data at hand that can possibly be developed

from the initial population of variables.

One of the greatest weaknesses of this method is that the initial population of variables must be selected on the basis of human judgment. If these variables are not selected on a purely logical basis, the remaining analysis can yield false or misleading results. Thus, great care and consideration must be exercised in the initial step of the process.

The method is limited in that the number of variables remaining in the study must be reduced to a maximum of nine in order to meet the limitations of the regression evaluation program. This problem can be circumvented by using a computer with a large memory capacity, thereby increasing the capacity of the program, or by analyzing the variables in groups of nine, and forming subgroups of the best variables from each group for further analysis. It is possible that the latter procedure might cause the elimination of a "good" variable by some accident of grouping. However, this occurrence seems highly unlikely, and probably need not be considered.

Use of this method, as developed here, must be tempered with a realization of the large amount of computer time involved in the final stages of analysis. The IBM 1620 computer was used for all work on this study. Approximately seven hours of computer time was involved in the study. The nature of the data used in the study, and the format required for input to the computer caused the data

input time to be the greatest time consuming factor. Actual calculation occupied a very small percentage of the time, but, due to memory limitations in the computer, all data had to be read in for each set of lead values considered. Since six sets of leads were used in this study, a deck of 240 data cards had to be read into the computer six times each for both the correlation program and the regression program. In both programs, the cards were read in pairs at about five second intervals. The long read cycle was the result of intermediate calculations made for each pair of data cards. Thus, unless a faster computer was available, the time consumed in an expanded study of this type would be prohibitive. Computer time could be reduced by consolidating the correlation and regression programs, reducing the number of sets of leads considered, or reducing the number of monthly observations included for each variable.

Despite the limitations mentioned, it is felt, based on the reasonably comprehensive nature of this study, that the power of this method of analysis has been verified, and that it should be considered if a demand forecast is to be attempted.

Extensions of the Study

The most fruitful area for study in the extension of this method is that of the initial selection of variables. Although it seems doubtful that an entirely objective

method could be developed for this purpose, it does seem possible that steps could be made in this direction. A suggested first step would be the development of a number of concisely stated criteria for inclusion or rejection of an examined variable in the study. Studies of this type generally rely on the intuition and judgment of the analyst. Formal criteria for selection seem to be nonexistent beyond the general statement of the need for logical relationships. Each analyst has his own series of informal and often unconscious criteria of selection, but this does not remove the selection process far enough from the realm of pure intuition.

If the methods of this study are selected for repeated use, an important addition would be the consolidation of the programs used into a single comprehensive program. Due to the nature of the programs, this would be a very difficult task, but if correctly done would greatly increase the efficiency of the method. The purpose of such an effort should primarily be the reduction of the computer time required. However, the addition of routines for computing residuals from the final regression equation, listing actual and forecast values for a selected number of periods, and testing for autocorrelation should be considered.

Further testing and evaluation with live data will, of course, contribute to the understanding of the usefulness and limitations of this technique.

Finally, comparison of this technique with other

forecasting methods using the same data would give measure to the relative strength of this particular method.

SELECTED BIBLIOGRAPHY

- (1) Bureau of Business Research. Oklahoma Business Bulletin. University of Oklahoma, 1953-1963.
- (2) Board of Governors of the Federal Reserve System. Federal Reserve Bulletin. Washington, D. C.
- (3) United States Department of Commerce. Business Statistics, Supplement to Survey of Current Business, 1957, 1991.
- (4) United States Department of Commerce. Survey of Current Business, 1962-63.
- (5) Ezekiel and Fox. Methods of Correlation and Regression Analysis. 3rd ed. New York: John Wiley and Sons, Inc., 1959.
- (6) International Business Machines Corporation. Statistical Programs. Section Six, IBM Users Library.
- (7) United States Department of Labor, Bureau of Labor Statistics. Monthly Labor Report.
- (8) Board of Governors of the Federal Reserve System. Industrial Production, 1957-59. Washington, D. C., 1963.
- (9) Wold, Herman. Demand Analysis. New York: John Wiley and Sons, 1953.
- (10) Peters and VanVoorhis. Statistical Procedures and Their Mathematical Bases. 1st ed. New York: McGraw-Hill Book Company, 1940.
- (11) DuBois. Multivariate Correlational Analysis. New York: Harper and Brothers, Publishers, 1957.
- (12) Spurr, Kellogg, and Smith. Business and Economic Statistics. Rev. ed. Homewood, Illinois: Richard D. Irwin, Inc., 1961.

- (13) Newbury, Frank D. Business Forecasting. 1st ed.
New York: McGraw-Hill Book Co., Inc., 1952.
- (14) Haney. Business Forecasting. New York: Ginn and
Company, 1931.
- (15) Self, Glendon D. Selection of Smoothing Constants
for an Exponentially Weighted Time Series
Model. (Ph.D. Dissertation, Oklahoma State
University), Aug., 1963.
- (16) Johnston, J. Econometric Methods. New York:
McGraw-Hill, 1960.

APPENDIXES

INTRODUCTION

Included in the following appendixes are all programs used in the preceding study. Included with each program is an explanation of its input format requirements, information pertinent to use of the program, a complete listing of the program, and the program output associated with the final regression equation developed in the study, if appropriate.

The program for calculation of correlation coefficients for all pairs of variables is based on a program found in the IBM Users Library, Program 6.0.052 (6), but was completely re-written to serve the needs of this study.

The program for evaluation of all regression equations for a given set of variables is a modification of a program from the IBM Users Library, Program 6.0.057. The program was modified to allow consideration of different sets of leads for the included variables. The portion of the program which developed and evaluated each regression equation was used without change.

The program for calculation of regression coefficients for any selected set of variables, is a supplement

to the above program. This program was used without modification.

Further, technical information on programs from the IBM Library can be obtained through reference to the original programs in the IBM Library catalogue (6).

PROGRAM I

Due to the selective nature of this procedure, variables are constantly being removed from the initial population. Input to the following programs requires that variables be sequentially arranged in the order in which they are called. Thus, as variables are eliminated, a new data deck must be punched omitting the eliminated variables. The following program was designed to select up to twelve variables, as specified by number, from a master variable deck including all 20 variables considered in this study. Input to the program is as follows:

The first card contains the numbers of the first eleven variables desired. The number of the first variable should appear in the first three columns of the card, the number of the second variable in the second three columns of the card, etc.

The second card should contain the number of the twelfth variable in the first three columns, and the number of variables to be punched out in the second three columns.

These two cards should be followed by the master data deck, with the first observation for each variable entered in consecutive ten column fields (use only

columns 1-70, i.e., seven variables per card) of the first three cards, the second observation on each variable entered in consecutive ten column fields of the next three cards, etc.

Output will be in the form of punched cards containing from seven to twelve variables, in the appropriate format for input to the following programs.

Program I

```

C   ARRANGES UP TO 12 VARIABLES INTO PROPER FORMAT FOR INPUT INTO
C   CORRELATION AND REGRESSION PROGRAMS--SELECTS FROM 20 VARIABLES
      DIMENSION V(21),N(12),T(12)
      READ 2,N(1),N(2),N(3),N(4),N(5),N(6),N(7),N(8),N(9),N(10),N(11)
2   FORMAT(13,13,13,13,13,13,13,13,13,13,13)
      READ 2,N(12),NVOUT
6   READ 4,V(1),V(2),V(3),V(4),V(5),V(6),V(7)
4   FORMAT(E10.4,E10.4,E10.4,E10.4,E10.4,E10.4,E10.4)
      READ 4,V(8),V(9),V(10),V(11),V(12),V(13),V(14)
      READ 4,V(15),V(16),V(17),V(18),V(19),V(20)
8   DO 5 I=1,12
      M=N(I)
      IF(M)1,18,5
5   T(I)=V(M)
18  PUNCH 4,T(1),T(2),T(3),T(4),T(5),T(6),T(7)
      IF(7-NVOUT)10,6,1
10  GO TO(1,1,1,1,1,1,1,1,11,12,13,14,15),I
11  PUNCH 4,T(8)
      GO TO 6
12  PUNCH 4,T(8),T(9)
      GO TO 6
13  PUNCH 4,T(8),T(9),T(10)
      GO TO 6
14  PUNCH 4,T(8),T(9),T(10),T(11)
      GO TO 6
15  PUNCH 4, T(8),T(9),T(10),T(11),T(12)
      GO TO 6
1   PRINT 16
16  FORMAT(33HLESS THAN 7 VAR, OR M IS NEGATIVE)
      END

```

SAMPLE INPUT DATA

37221.	405598.	95043.	128493.	89.0	100.5	72.0
102.5	98.3	88.5	117.2	85.5	812.	3583.
0.0	0.0	0.0	120.0	74437.	12805.	
50752.	396145.	98187.	128517.	109.4	116.7	84.0
111.1	103.1	96.8	120.6	112.9	-12620.	-01317.
0.0	0.0	0.0	119.0	87539.	17657.	
59340.	411639.	116553.	126088.	113.6	124.6	93.4
116.7	102.5	100.8	122.5	126.4	-14210.	-499.
0.0	0.0	0.0	118.0	88789.	16790.	

SAMPLE OUTPUT DATA

.3722E 05	.1284E 06	.7200E 02	.8550E 02	.0000E-99	.0000E-99	.0000E-99	#001
.1200E 03	.7443E 05						#002
.5075E 05	.1285E 06	.8400E 02	.1129E 03	.0000E-99	.0000E-99	.0000E-99	#003
.1190E 03	.8753E 05						#004
.5934E 05	.1260E 06	.9340E 02	.1264E 03	.0000E-99	.0000E-99	.0000E-99	#005
.1180E 03	.8878E 05						#006

PROGRAM II

The following program will compute the simple correlation coefficients for all pairs of up to twelve variables. The program is designed to make calculations for any set of lead values corresponding to the input variables. Lead values may be zero, but must be less than fifteen. The correlation coefficients are computed from the equation

$$r_{ij} = \frac{N \sum X_i X_j - \sum X_i \sum X_j}{\sqrt{N \sum X_i^2 - (\sum X_i)^2} \sqrt{N \sum X_j^2 - (\sum X_j)^2}} .$$

Input to the program should consist of the following:

The first card contains the number of variables in columns one to five, the number of observations on each variable in columns six to ten, the maximum lead value plus one in columns eleven to thirteen, and a three digit identification for the set of lead values under consideration in columns fourteen to sixteen.

The second card contains the number of periods lead for the first variable, in columns one to three.

Each succeeding card contains the lead value for consecutive variables, until a lead value has been

provided for each input variable. Each of these values occupies columns one to three on separate cards.

The data deck punched by Program I, or a data deck in the format described for the master data deck, modified for the appropriate number of variables, should follow.

The last card should contain a one in column 72.

Output from the program is printed on the typewriter, and consists of the number of variables, the number of observations on each variable, the identification of the set of lead values used, and the i - j subscripts for each pair of variables followed by the correlation coefficient for that pair of variables, as shown in the sample following the program listing.

Program II

```

C      SIMPLE CORRELATION FOR ALL PAIRS--UP TO 12 VARIABLES
      DIMENSION T(12,15),N(12),SUM(12),SUM2(12),PROD(12,12),X(12)
      DIMENSION R(12,12)
1      READ 20,N1,N2,M,NL
20     FORMAT(15,15,13,13)
      N3=N1-1
      Y2=N2
      DO 5 I=1,N1
      SUM(I)=0.
      SUM2(I)=0.
      DO 5 J=1,N1
5      PROD(I,J)=0.
      DO 21 I=1,N1
21     READ 22, N(I)
22     FORMAT(13)
      DO 26 J=1,M
23     READ 24,T(1,J),T(2,J),T(3,J),T(4,J),T(5,J),T(6,J),T(7,J),J1
24     FORMAT(E10.4,E10.4,E10.4,E10.4,E10.4,E10.4,E10.4,12)
      IF(J1)33,25,33
25     IF(7-N1)26,27,27
26     READ 24,T(8,J),T(9,J),T(10,J),T(11,J),T(12,J)
27     DO 28 I=1,N1
      L=N(I)+1
28     X(I)=T(I,L)
      DO 30 I=1,N1
      DO 30 J=1,M
30     T(I,J)=T(I,J+1)
      DO 50 I=1,N1
      SUM(I)=SUM(I)+X(I)
50     SUM2(I)=SUM2(I)+X(I)*X(I)
      DO 7 J=1,N3
      L=J+1
      DO 7 K=L,N1
7      PROD(J,K)=PROD(J,K)+X(J)*X(K)
      J=M
      GO TO 23
33     DO 8 I=1,N3
      L=I+1
      DO 8 J=L,N1
      XNUM=(Y2*PROD(I,J)-SUM(I)*SUM(J))
      DEN=(SQR((Y2*SUM2(I)-SUM(I)*SUM(I))*(Y2*SUM2(J)-SUM(J)*SUM(J))))
8      R(I,J)=XNUM/DEN
      DO 9 I=1,N1
9      R(I,I)=1.0
      PRINT 20,N1,N2,NL
      PRINT 35
35     FORMAT(/24H I J R(I,J)/)
      DO 10 I=1,N1
      DO 10 J=1,N1
10     PRINT 32,I,J,R(I,J)
32     FORMAT(15,15,5X,E14.8)
      PAUSE
      GO TO 1
      END

```

Program II (Continued)

12 120 6

I	J	R(I,J)	I	J	R(I,J)
1	1	.10000000E+01	6	8	.94129175E-00
1	2	.55805458E-00	6	9	.72246251E-00
1	3	.43559411E-00	6	10	.70991747E-00
1	4	.37588722E-00	6	11	.92256220E-00
1	5	.56609331E-00	6	12	.90025174E-00
1	6	.62606318E-00	7	7	.10000000E+01
1	7	.21144306E-00	7	8	.70545130E-00
1	8	.58350755E-00	7	9	.58647057E-00
1	9	.53001619E-00	7	10	.70734489E-00
1	10	.47166657E-00	7	11	.67190584E-00
1	11	.57342606E-00	7	12	.47125878E-00
1	12	.68434757E-00	8	8	.10000000E+01
2	2	.10000000E+01	8	9	.82532573E-00
2	3	.77392675E-00	8	10	.87957696E-00
2	4	.75464145E-00	8	11	.96725561E-00
2	5	.74179982E-00	8	12	.88003153E-00
2	6	.94377816E-00	9	9	.10000000E+01
2	7	.52387031E-00	9	10	.83201266E-00
2	8	.91072258E-00	9	11	.82964202E-00
2	9	.68011159E-00	9	12	.73824600E-00
2	10	.65630116E-00	10	10	.10000000E+01
2	11	.92860641E-00	10	11	.81397254E-00
2	12	.79118822E-00	10	12	.75154912E-00
3	3	.10000000E+01	11	11	.10000000E+01
3	4	.85642180E-00	11	12	.80717300E-00
3	5	.44768663E-00	12	12	.10000000E+01
3	6	.72728112E-00			
3	7	.26184722E-00			
3	8	.64381196E-00			
3	9	.37812467E-00			
3	10	.32529565E-00			
3	11	.67040328E-00			
3	12	.56420333E-00			
4	4	.10000000E+01			
4	5	.42460151E-00			
4	6	.74073318E-00			
4	7	.18756832E-00			
4	8	.61199080E-00			
4	9	.29835029E-00			
4	10	.25894613E-00			
4	11	.61634765E-00			
4	12	.57504840E-00			
5	5	.10000000E+01			
5	6	.83487132E-00			
5	7	.62762242E-00			
5	8	.92120082E-00			
5	9	.82615872E-00			
5	10	.91961172E-00			
5	11	.83278288E-00			
5	12	.88462328E-00			
6	6	.10000000E+01			
6	7	.52024923E-00			

PROGRAM III

This program computes the coefficient of determination and the residual sum of squares for all regression equations that can be formulated from up to nine input variables. The program accepts data on the basis of lead values, in the same manner as described for the previous program. The coefficient of determination for each regression equation is compared with a specified test value, and a card is punched for each equation whose coefficient of determination exceeds this value. This card contains the number of variables in the combination, the coefficient of determination, the residual sum of squares, and a listing of the variables included in the combination. If program switch 1 on the 1620 is off during processing, these cards follow an identification card. If switch 1 is on, the triangular matrix of sums of cross products is first punched. These cards serve as input to the next program, and will be explained later.

Input to the program should be in the following form:

The first fifteen columns of the first card may be used for identification, and will be reproduced in the output identification card. Columns 19 and 20 contain the number of variables; columns 22 to 25

contain a three digit test value of the form .XXX; column 30 should contain a zero if it is desired to include the constant term in the equation, and a one if the constant term is to be omitted; columns 31 to 33 contain a three digit identification of the set of lead values being used; and columns 34 to 36 contain the maximum lead value plus one. If the constant term is omitted from the equations, ten variables may be included instead of nine.

The next n cards each contain a single lead value in columns 1 to 3, for each of the n variables.

These cards are followed by a data deck as punched by program one, or in the format described for the master data deck. The dependent variable must be the last one entered, for the program develops regressions against this last variable, whatever it may be.

The last card must contain a one in column 72.

Data must be re-entered for each set of lead values analyzed.

Program output has been previously described, and is shown in the sample following the program listing.

Program III

```

C   ALL POSSIBLE REGRESSIONS FOR UP TO 12 VARIABLES
C   SWITCH 1 ON WILL PUNCH MATRIX OF SUMS OF CROSS PRODUCTS

      DIMENSION A(220),M(9),X(11),G(11,15),LEAD(11)
      LIM=10
      DO 1 I=1,9
1     M(I)=0
      X(1)=1.0
2     READ 3,NVAR,TEST,J2,NLEAD,MLEAD
3     FORMAT(15H                15,F5.3,15,13,13)
      IF(J2)4,5,4
4     J2=1
5     MS=1-J2
      MSIZE=NVAR+MS
      IF(LIM-MSIZE)6,8,8
6     J=LIM-MS
      PRINT 7,NVAR,J
7     FORMAT(15,20H VARIABLES, LIMIT IS 13)
      PAUSE
      GO TO 2
8     INC=MSIZE*(MSIZE+1)/2
      DO 9 I=1,INC
9     A(I)=0.0
      I1=2-MS
      I2=NVAR+1
      N=0

C   READ IN SAMPLES AND BUILD UP TRIANGULAR MATRIX

      DO 100 I=2,12
100    READ 101,LEAD(I)
101    FORMAT(13)
      PRINT 106,NLEAD
106    FORMAT(13HLEAD DECK ID 13)
      DO 13 J=1,MLEAD
11     FORMAT(E10.4,E10.4,E10.4,E10.4,E10.4,E10.4,E10.4,12)
10     READ 11,G(2,J),G(3,J),G(4,J),G(5,J),G(6,J),G(7,J),G(8,J),J1
      IF(J1)17,12,17
12     IF(7-NVAR)13,105,105
13     READ 11,G(9,J),G(10,J)
105    DO 107 I=2,12
      L=LEAD(I)+1
107    X(I)=G(I,L)
      DO 103 I=2,12
      DO 103 J=1,MLEAD
103    G(I,J)=G(I,J+1)
15     K=0
      DO 16 I=1,12
      P=X(I)
      DO 16 J=1,12
      K=K+1
16     A(K)=A(K)+P*X(J)

```

Program III (Continued)

```

      N=N+1
      J=MLEAD
      GO TO 10
17  PUNCH 3,NVAR,TEST,J2
      PUNCH 18,N,A(INC)
18  FORMAT(11HNO SAMPLES=15,13H, RAW S OF S=E11.4)
      IF(SENSE SWITCH 1)19,22

C    PUNCH MATRIX OF SUMS OF PRODUCTS

19  K=0
      DO 20 I=J2,NVAR
      DO 20 J=I,NVAR
      K=K+1
20  PUNCH 21,I,J,A(K)
21  FORMAT(115,15,E20.8)

C    SET UP AND SOLVE REGRESSIONS

22  T=A(INC)
      S=1.0/T
      K1=1
      KP=1
      I1=2
      KS=MSIZE
      PUNCH 23
23  FORMAT(/39HNVAR   R-SQ   RES S OF S   VARIABLE LIST)
24  K2=MSIZE-KP
      L1=KP-MS-1
      DO 31 K=K1,K2
      INC=INC-KS
      KS=KS-1
      I2=I1+KS-KP
      J1=I2+1
      P=1.0/A(I1-1)
      DO 26 I=I1,I2
      Q=P*A(I)
      L=J1-I
      J2=L+I2
      DO 25 J=J1,J2
      JK=J-L
      JN=J+INC
25  A(JN)=A(J)-Q*A(JK)
26  J1=J2+1
      M(K)=K+L1
      IF(K+L1)6,27,28
27  T=A(JN)
      S=1.0/T
28  P=A(JN)
      Q=(T-P)*S
      IF(Q-TEST)31,29,29
29  N=K-MS
      I=M(9)
      PUNCH 30,N,Q,P,M(1),M(2),M(3),M(4),M(5),M(6),M(7),M(8),I
30  FORMAT(13,F8.4,E12.4,16,13,13,13,13,13,13,13,13)
31  I1=I2+INC+2

```


Program III (Continued)

```
M(K2)=0
K1=K2-1
IF(K1-MS)6,2,32
32 KP=M(K1)+MS+2-K1
KS=KS+2
INC=INC+KS+KS-1
I1=J1-INC-INC+KP*(KS+KS+3-KP)/2
GO TO 24
END
```

Program III (Continued)

SALES DATA 9 .464 0
 NO SAMPLES= 116, RAW S OF S= 4.8304E+11

NVAR	R-SQ	RES S OF S	VARIABLE LIST											
8	.4704	26.4623E+09	0	1	2	3	4	5	6	7	8			
7	.4695	26.5075E+09	0	1	2	3	4	5	7	8	0			
7	.4703	26.4656E+09	0	1	2	4	5	6	7	8	0			
6	.4694	26.5126E+09	0	1	2	4	5	7	8	0	0			
7	.4695	26.5036E+09	0	2	3	4	5	6	7	8	0			
6	.4688	26.5428E+09	0	2	3	4	5	7	8	0	0			
6	.4695	26.5036E+09	0	2	4	5	6	7	8	0	0			
5	.4687	26.5435E+09	0	2	4	5	7	8	0	0	0			

PROGRAM IV

This program is an extension to the previous program, and is taken without change from the IBM Users Library. Complete information on this program may be obtained from the Users Library catalogue. This program computes regression coefficients for selected combinations of variables.

Input/output formats and a program listing are included here only in the interest of completeness.

Input to this program consists of selected output from Program III as follows:

If this program is entered immediately following Program III, and program switch 1 is off, enter the identification card (the first card punched as output from Program III) and any variable cards, from Program III, for which regression coefficients are desired.

If switch 1 is on, enter the identification card, the matrix of sums of cross products cards, and variable cards for which regression coefficients are desired.

If program switch one is off, the program assumes that the matrix of cross products is already in memory. Thus, if it is impossible to enter this program immediately

following Program III, the cross products matrix cards must be obtained from Program III for input to this program.

Output consists of the identification card, the residual sum of squares for the regression equation, the multiple correlation coefficient, and the regression coefficients for each variable, as shown in the sample following the program listing.

Program IV

```

C      EXTENSION TO PROGRAM FOR REGRESSION ON COMBINATIONS OF VARIABLES
C      TO FIND REGRESSION COEFFICIENTS OF SELECTED COMBINATIONS
C      BY E.VERNON GRIFFITH - IBM
C      SWITCH 1 ON AFTER HEADER CARD PROGRAM WILL READ IN MATRIX CARDS
C      OFF PROGRAM ASSUMES MATRIX ALREADY IN MEMORY
C      SWITCH 2 ON AFTER SOLUTION WILL RETURN TO READ NEW HEADER CARD
C      OFF AFTER SOLUTION WILL READ IN NEW COMBINATION

      DIMENSION A(105),M(14),B(13,14)
1 READ 2,NVAR,TEST,J2
2 FORMAT(15H          15,F5.3,15)
  PUNCH 2,NVAR,TEST,J2
  MS=1-J2
  MSIZE=NVAR+MS
  IF(SENSE SWITCH 1)3,9
C READ IN MATRIX
3 K=0
  DO 8 I=J2,NVAR
  DO 8 J=1,NVAR
  K=K+1
  READ 4,I1,J1,A(K)
4 FORMAT(115,15,E20.8)
  IF(I-I1)6,5,6
5 IF(J-J1)6,8,6
6 PRINT 7,I,J
7 FORMAT(26HINPUT ERROR-SUBS SHOULD BE 13,13)
  PAUSE
  GO TO 1
8 CONTINUE

C READ IN COMBINATION
9 READ10,N,Q,P,M(1),M(2),M(3),M(4),M(5),M(6),M(7),M(8),M(9),I,J,L,11
10 FORMAT(13,F8.4,E12.4,16,13,13,13,13,13,13,13,13,13,13,13)
  M(10)=I
  M(11)=J
  M(12)=L
  M(13)=11
  L=N+MS
  M(L+1)=NVAR
  Q=SQR(Q)
  PUNCH 11,P,Q
11 FORMAT(/13HRESID S OF S=E11.4,12H, MULT CORR=F7.4)

C SELECT MATRIX ELEMENTS
  K=1
  I1=1
  I2=NVAR-1
  DO 16 I=J2,12
  IF(M(I1)-I)6,13,12
12 K=K+NVAR-I+1
  GO TO 16
13 J1=11
  DO 15 J=1,NVAR

```

Program IV (Continued)

```

      IF(M(J1)-J)6,14,15
14  B(I1,J1)=A(K)
      J1=J1+1
15  K=K+1
      I1=I1+1
16  CONTINUE

```

C SOLVE FOR REGRESSION COEFFICIENTS AND OUTPUT

```

      DO 25 K=1,L
      P=1.0/B(K,K)
      J1=K+1
      DO 18 J=J1,I1
      B(J,K)=B(K,J)
18  B(K,J)=B(K,J)*P
      DO 25 I=1,L
      IF(I-K)20,22,19
19  J1=I
20  P=B(I,K)
      DO 21 J=J1,I1
21  B(I,J)=B(I,J)-B(K,J)*P
22  IF(L-K) 6,23,25
23  PUNCH 24,M(I),B(I,I1)
24  FORMAT(9H  VAR NO 13,7H  COEF=E11.4)
25  CONTINUE
      IF(SENSE SWITCH 2) 1,9
      END

```

Program IV (Continued)

SALES DATA 9 .464 0

RESID S OF S= 2.6462E+10, MULT CORR= .6858

VAR NO	0	COEF=	7.1432E+04
VAR NO	1	COEF=	4.4068E-02
VAR NO	2	COEF=	2.3439E-01
VAR NO	3	COEF=	7.2842E-00
VAR NO	4	COEF=	-3.9979E+02
VAR NO	5	COEF=	-9.5335E+03
VAR NO	6	COEF=	2.9038E+03
VAR NO	7	COEF=	6.7358E+03
VAR NO	8	COEF=	2.2176E+02

RESID S OF S= 2.6507E+10, MULT CORR= .6852

VAR NO	0	COEF=	7.6649E+04
VAR NO	1	COEF=	4.0713E-02
VAR NO	2	COEF=	2.4086E-01
VAR NO	3	COEF=	9.1529E-00
VAR NO	4	COEF=	-4.4681E+02
VAR NO	5	COEF=	-1.0670E+04
VAR NO	7	COEF=	5.8924E+03
VAR NO	8	COEF=	2.2511E+02

RESID S OF S= 2.6465E+10, MULT CORR= .6857

VAR NO	0	COEF=	7.2290E+04
VAR NO	1	COEF=	4.1107E-02
VAR NO	2	COEF=	2.3329E-01
VAR NO	4	COEF=	-3.9975E+02
VAR NO	5	COEF=	-9.5223E+03
VAR NO	6	COEF=	2.9576E+03
VAR NO	7	COEF=	6.8259E+03
VAR NO	8	COEF=	2.2207E+02

VITA

Douglas Lathel Johnson

Candidate for the Degree of
Master of Science

Thesis: FORECASTING SALES FOR THE SMALL READY-MIX
CONCRETE COMPANY

Major Field: Industrial Engineering and Management

Biographical:

Personal Data: Born in Borger, Texas, June 24, 1940,
the son of Lathel M. and Virginia L. Johnson.

Education: Attended grade school in Phillips, Texas,
and Bartlesville, Oklahoma; graduated from
College High School in 1958; received the
Bachelor of Science degree from Oklahoma State
University, with a major in Industrial Engineer-
ing and Management, in January, 1963; completed
requirements for the Master of Science degree in
August, 1963.

Professional Experience: Employed by Eastman Kodak
Company as an engineering trainee during the
Summer of 1962. Employed by Phillips Petroleum
Company as an engineering trainee during the
Summers of 1961 and 1960, and as a laborer dur-
ing the Summer of 1959. Served six months active
duty with the Army Reserve from June, 1958 to
January, 1959.

College Activities: Member of Alpha Pi Mu (honorary
Industrial Engineering fraternity), member of
American Institute of Industrial Engineers, and
member of Sigma Tau (honorary Engineering
fraternity).